Conservation of expression regulation throughout the animal kingdom

Michael Kuhn and Andreas Beyer

bioRxiv first posted online July 18, 2014
Access the most recent version at doi: http://dx.doi.org/10.1101/007252

Conservation of expression regulation throughout the animal kingdom

Michael Kuhn[1,2*], Andreas Beyer[3*]

[1] Biotechnology Center, TU Dresden, Dresden, Germany

[2] Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany

[3] University of Cologne, Cologne, Germany

* To whom correspondence should be addressed: michael.kuhn@biotec.tu-dresden.de (Phone: +49-351-463 40064, Fax: +49-351-463 40061), andreas.beyer@uni-koeln.de (Phone: +49 221-478 84429, Fax: +49 221-478 84045)

Gene expression programs have been found to be highly conserved between closely related species, especially when comparing the same tissue types between species. Such analysis is, however, much more challenging over larger evolutionary distances when complementary tissues cannot readily be defined. Here, we present the first cross-species mapping of tissue-specific and developmental gene expression patterns across a wide range of animals, including many non-model species. Importantly, our approach does not require the definition of homologous tissues. In our survey of 32 datasets across 23 species, we detected conserved expression programs on all taxonomic levels, both within animals and between the animals and their closest unicellular relatives, the choanoflagellates. We found that the rate of change in tissue expression patterns is a property of gene families. Subsequently, we used the conservation of expression programs as a means to identify neofunctionalization of gene duplication products. We found 1206 duplication events where one of the two genes kept the expression program of the original gene, whereas the other copy adopted a novel expression program. We corroborated such potential neofunctionalizations using independent network information: the duplication product with the more conserved expression pattern shared more interaction partners with the non-duplicated reference gene than the more divergent duplication product. Our findings open new avenues of study for the comparison and transfer of knowledge between different species.

Introduction

Gene functions have traditionally been determined using molecular and cellular approaches involving forward or reverse genetics. Functional annotations that were directly determined through these approaches are, however, not available at all for most species, and incomplete even for model species (Thomas et al. 2012). For non-model species, often only data transferred from other organisms is available. In this case, the degree of conservation of functions is uncertain, especially when a gene is duplicated in a non-model species, but not in the model species where its function has originally been studied. Previously, gene co-expression data has been used to find conserved co-expressed modules (Stuart 2003; Gerstein et al. 2014) and to uncover functional similarities between genes from different species (Chikina & Troyanskaya 2011). However, the latter approach requires that the two species are well-studied in both gene expression and functional annotation, and will suffer from incomplete and biased annotations (Thomas et al. 2012). Developmental gene expression profiles between closely related species can be compared to find functional links between genes and to detect differences between orthologs (Yanai et al. 2011; Levin et al. 2012; Silver et al. 2012). Existing approaches require that expression datasets have been obtained under comparable conditions for the respective species. For closely related species, homologous tissues can easily be identified (Niknejad et al. 2012), and cross-species correlations between homologous tissues of closely related species have previously been investigated (Piasecka, Kutalik, et al. 2012a; Liao & Zhang 2006). This is however a severe limitation for functional mapping between many species. Even between closely related species, the relative amounts of cell types that make up tissues may change. Across larger evolutionary distances, only few clearly homologous tissues are available. Nonetheless, it is possible to identify deep homologies among tissues (Shubin et al. 2009). For example, homologous structures have been identified in the nervous systems of vertebrates and annelids (Tomer et al. 2010; Strausfeld & Hirth 2013). Other organs show functional convergence, for example mammalian liver and brown fat in flies, which both carry out xenobiotic clearance functions (Chung et al. 2009).

Many gene expression datasets have been generated under experimental conditions that represent non-physiological conditions, such as gene knockouts, which are not under evolutionary selection. Such data is therefore not necessarily

suitable for comparing gene expression across species (Seok et al. 2013). In contrast, the formation of tissues during development and the maintenance of tissue function throughout the life of an animal are crucial for survival and reproduction, and are therefore under direct evolutionary selection (Winter et al. 2004; Gu & Z. Su 2007). Tissue expression data is available for many species, as tissues can be gathered even from non-model species where genetic tools such as transgenesis or RNAi are not available. Previous research has shown that it is possible to predict tissue-specific expression patterns from gene expression experiments within the same species (Chikina et al. 2009). However, it remains challenging to map tissue expression over larger phylogenetic distances. If such mapping was possible, we could substantially improve the annotation of non-model-species genomes, fill annotation gaps in model species, and in particular address the problem of gene duplications.

We have developed a method to map tissue expression patterns of genes from one species to another, without defining equivalent tissues between the two species. For each gene of the source species, this approach predicts a virtual tissue expression pattern in the destination species. These virtual expression patterns can then be compared to the expression patterns of genes in the target species, enabling us to calculate correlations between the expression patterns of genes across species. We showed that high correlations in tissue expression across species are predictive for 1:1 orthology, shared structure, and similar function. Subsequently we used our modeling approach for three applications: first, for determining the degree of conservation of tissue-specific gene expression patterns, second, for comparing the speed of functional divergence between independently evolving members of protein families, and third, for analyzing the fate of gene duplication products.

Results

Correlation between tissues of distant species

To analyze tissue expression across the entire metazoan kingdom, we gathered genome and tissue expression data from 32 datasets covering 23 different species (Table S1). Among these were eight chordate species: *Ciona intestinales* (Shoguchi et al. 2011), *Danio rerio* (Domazet-Loso & Tautz 2010), *Gallus gallus* (Chan et al. 2009; Irie & Kuratani 2011), *Homo sapiens* (Lukk et al. 2010), *Mus musculus* (Irie & Kuratani 2011; A. I. Su et al. 2004), *Sus scrofa* (Freeman et al. 2012), *Tetraodon*

nigroviridis (Chan et al. 2009), and *Xenopus tropicalis* (Chan et al. 2009; Yanai et al. 2011); two cnidarians: *Hydra vulgaris* (Hemmrich et al. 2012) and *Nematostella vectensis* (Tulin et al. 2013); two flatworms: *Schistosoma japonicum* (Gobert et al. 2009) and *Schistosoma mansoni* (Nawaratna et al. 2011; Fitzpatrick et al. 2009); three insects: *Anopheles gambiae* (Baker et al. 2011; Dissanayake et al. 2006; Goltsev et al. 2009), *Bombyx mori* (Xia et al. 2007) and *Drosophila melanogaster* (St Pierre et al. 2014; Robinson et al. 2013); seven nematodes: *Ascaris suum* (Wang et al. 2013), *Brugia malayi* and five *Caenorhabditis* species (Levin et al. 2012; Spencer et al. 2011). Furthermore, we added the choanoflagellate *Salpingoeca rosetta* as an outgroup (Fairclough et al. 2013). Many datasets contain both tissues and developmental samples, e.g. different adult organs and embryonic stages. For the sake of brevity, we refer to the all of these samples as “tissues.”

To determine orthology relations between genes, we assembled groups of orthologs (OGs) using the eggNOG pipeline (Powell et al. 2014) on the genomes of the choanoflagellate *Salpingoeca rosetta* and 67 animals. We then computed gene trees for all OGs using GIGA (Thomas 2010), which we then analyzed to extract 1:1 orthologs and duplication events. First, we quantified the correlation of gene expression between tissues across species. For each pair of datasets, we built gene expression vectors for all tissues using the expression patterns of 1:1 orthologs. This yielded one vector of expression values for each tissue. We then calculated the correlation of these vectors across species and found that for 89.0% of all dataset pairs, more than half of the tissues in one dataset were significantly correlated with at least one tissue from the other dataset (using a p-value cutoff of 0.05 for each tissue pair). Importantly, this was true even across large phylogenetic distances. For example, between fly and *C. elegans*, the two largest correlations of 0.31 were between ovary and gonad, and between head and L2 glutamate receptor neurons. When we removed the three worst datasets from the analysis (*Nematostella, Hydra* and *Bombyx*), the fraction increased to 98.4% of all dataset pairs. Interestingly, for 70.6% of all dataset pairs in the filtered set, all tissues of one dataset were significantly correlated with at least one tissue from the second dataset. These correlations suggested that it is feasible to map gene expression patterns between tissues of distantly related species, even if a homology relation between the tissues is not apparent.

Mapping gene expression between species

To predict tissue expression patterns across species we chose a simple and transparent method, namely to train linear models for mapping expression values across species (Fig. 1). Given the tissue expression values for a source species, each linear model predicted the expression value for one tissue from the target species. Thus, for each combination of source and target species, we trained as many linear models as there are tissues in the target species. Importantly, this modeling approach did not require 1:1 relationships of tissues (i.e. the existence of homologous tissues). Rather, the expression in each tissue of the target species was modeled as a combination of the tissues in the source species (see Methods).

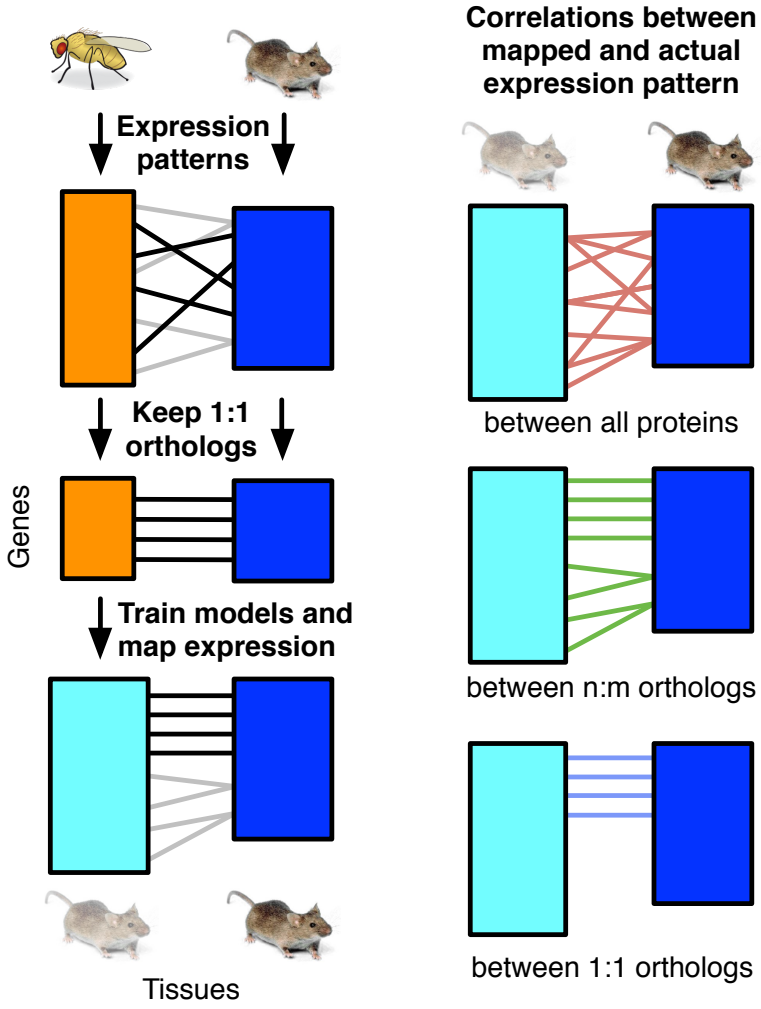


Fig. 1: Mapping expression patterns across species. For each tissue in the target species, models were trained to predict the tissue-specific gene expression pattern from the expression patterns of 1:1 orthologs in a source species. Mapping the expression patterns of all genes created virtual expression patterns, which could then be used to compute correlations between the mapped and actual expression patterns.

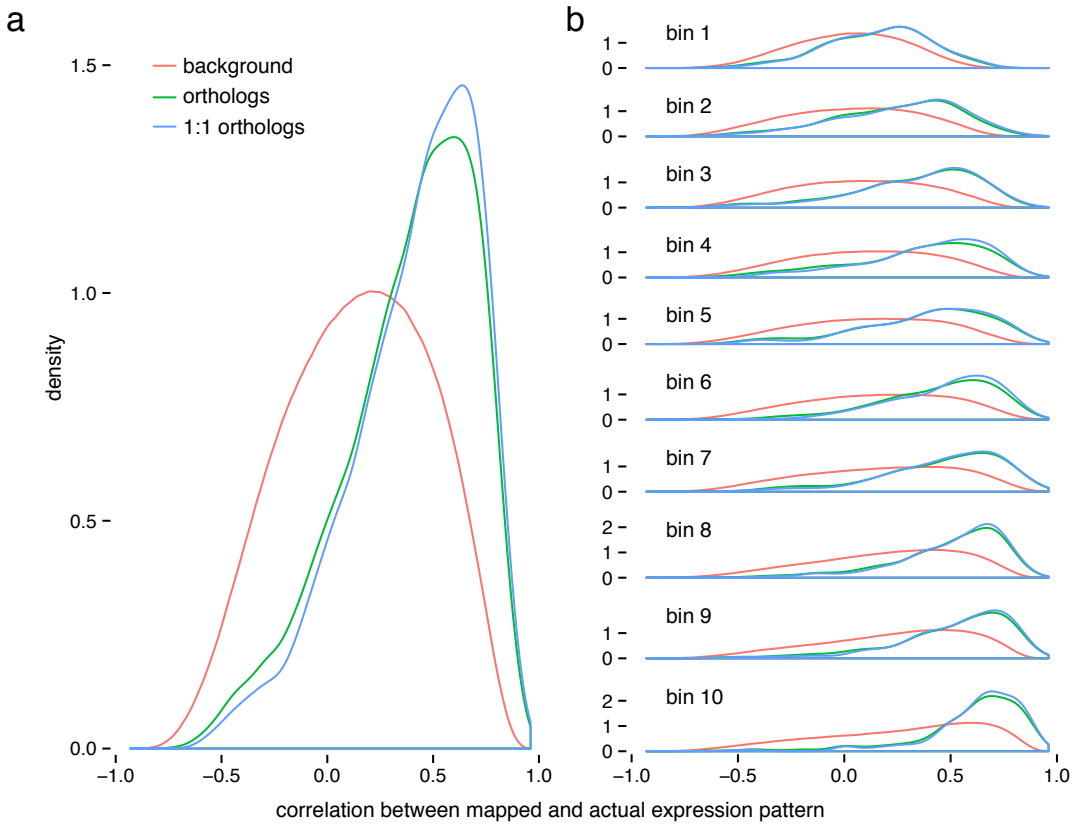


Fig. 2. Distribution of correlations between mapped and actual expression patterns. (a) When mapping expression patterns from fly to *C. elegans*, correlations between orthologs (green) and 1:1 orthologs (blue) were much higher than for background gene pairs (pairs of genes that are not homologous to each other, shown in red). **(b)** Target genes were distributed in bins according to the number of genes with similar expression patterns within the target species. Pairs of background genes had a higher correlation when there were more genes with similar expression patterns, as is evident from the shift towards higher correlations. For this pair of datasets, bins contained between 297 and 316 one-to-one orthologs, with an average of 305.

Using the trained model, we mapped all expression values from the source to the target species. We then calculated the Pearson correlation between the mapped expression values and the actual expression values, for three sets of genes: (1) all genes having homologs between the two species, (2) orthologous groups and (3) 1:1 orthologs. We restricted the background set (group 1) to genes with homologs to exclude lineage-specific genes that were found to have much lower correlations than genes with homologs. When analyzing a pair of genes that are 1:1 orthologs, we used expression values predicted by 10-fold cross-validation. From the distribution of correlations, we calculated p-values for all pairs of genes using the null hypothesis that the compared genes belong to the background and thus are not orthologous. During initial tests, we found a strong correlation between these p-values and the number of genes with similar expression patterns in the target

species (Fig. S1, dashed lines). We therefore split target genes into bins according to the number of target genes with similar expression patterns (Fig, 2b). For each bin, we obtain a mapping from correlation to p-values. For a given correlation between the mapped expression pattern of the source gene and the expression pattern of the target gene, we then calculate an expression distance out of the p-values obtained for the adjacent bins (see Methods). Thus, a low expression distance indicates that the expression of this gene in a given target species can be well predicted using the expression of homologous genes in the source species.

The mapping success can be measured in different ways. For each pair of datasets, we first compared the distribution of correlations for background genes and 1:1 orthologs using the Kolmogorov-Smirnov (K-S) test. Controlling for multiple testing with the Benjamini-Hochberg method (Benjamini & Hochberg 1995), 77% of all K-S p-values were significant ($q < 0.05$). As K-S p-values are strongly influenced by the number of gene pairs, we also computed the fraction of 1:1 orthologs that can be mapped at an expression distance threshold of 0.25 (Fig. 3). This fraction was highly correlated with the K-S statistic *D* (Pearson correlation 0.97), but more intuitive. Across all pairs of datasets, the median fraction of 1:1 orthologs with expression distances below 0.25 was 40%, indicating an enrichment of 1:1 orthologs with well-conserved expression patterns. This analysis revealed both an expected enrichment for closely related species and unexpectedly high enrichments between very distant species, such as between chordates and insects. In general, developmental datasets mapped less well to other species than datasets of adult tissues. However, this difference could be attributed to the information content of the different datasets, which we did not quantify.

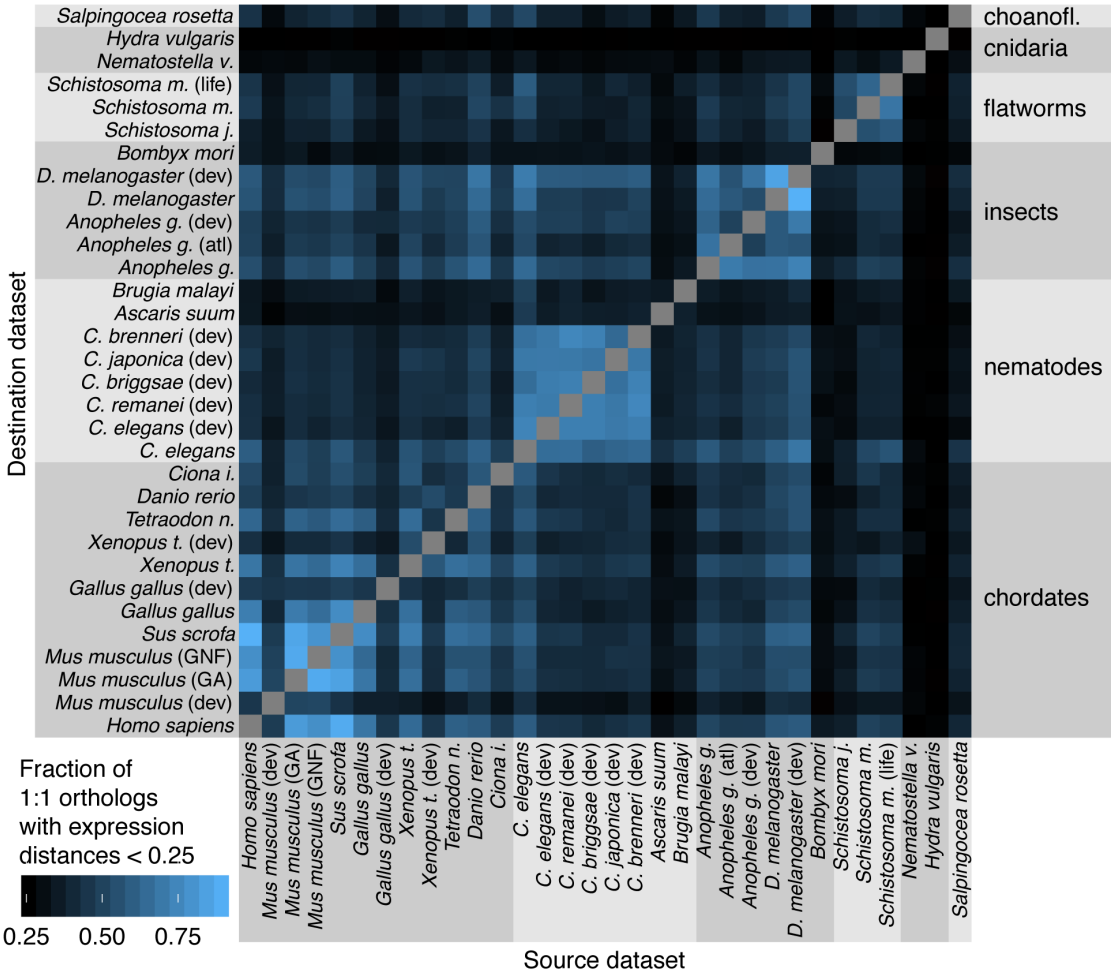


Fig. 3. Conservation of expression patterns throughout the metazoans and choanoflagellates. For all dataset pairs, the fraction of 1:1 orthologs with expression distances below 0.25 is shown. Within clades, this fraction becomes very high and approaches 1 in some cases. When there is no enrichment of 1:1 orthologs towards lower expression distances, the distributions of correlations are identical for pairs of background genes and orthologs. In this case, the distribution of expression distances is uniform, and the fraction of orthologs with expression distances below 0.25 is 0.25 (see Fig. 5). Note that there are some datasets with universally low values. Here, the kinds of measured tissues and the quality of the dataset apparently prevented better mapping performance. However, some otherwise distant species had a higher than expected fraction of 1:1 orthologs with well-conserved expression patterns.

Benchmarks

In order to establish the biological relevance of our expression distance measure, we applied benchmarks at three levels, namely sequence, structure, and function. On the sequence level, we found that expression distances could be used to decide which of the top two BLAST hits for a query protein is the true 1:1 ortholog of the query protein in the target species (Fig. 4A and Fig. S3). On the structural level (Lees et al. 2014), expression distance and the number of proteins belonging to a structural fold were correlated (Fig. 4B and Fig. S4). That is, structural folds with

fewer members, and hence lower functional diversity, were more similar in their expression patterns across species. Lastly, on the functional level, we applied the phenolog concept (McGary et al. 2010) to find equivalent phenotypic annotations across species. We found that expression distances could be used to predict which member of a protein family has been annotated with a phenotype (Fig. 4C and Fig. S5).

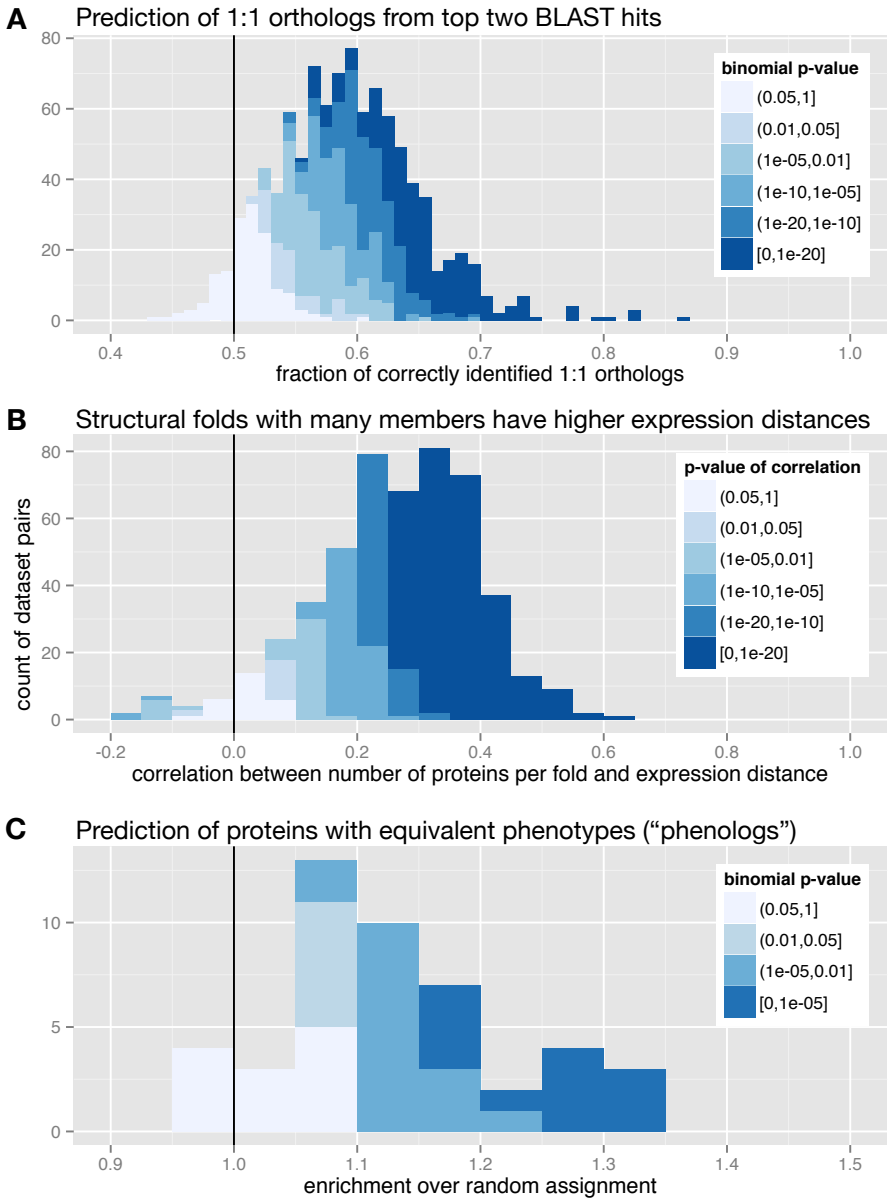


Fig. 4. Summary of benchmarking results. For each pair of datasets from different species, the performance in different benchmarks has been computed, along with a p-value. In all three benchmarks, there was a clear shift of the results relative to the random expectation (black line). For details, see Fig. S3, Fig. S4, Fig. S5 and supplementary text. Due to limited structural and functional annotations, there was a lower number of dataset pairs for the two lower panels.

Conservation of gene expression programs

At all taxonomic levels, we determined the conservation of the expression patterns of 1:1 orthologs. This data then allowed us to estimate the degree of conservation of tissue-specific expression patterns, even between groups of species that do not have readily identifiable homologous organs. For all sets of 1:1 orthologs, we computed the median expression distances when mapping across a particular taxonomic split (e.g. for vertebrates, we mapped between fish and tetrapods). First, we compared the distributions of expression distances to the uniform distribution. With the exception of mappings with cnidaria (*Nematostella* and *Hydra*), all distributions were significantly shifted towards lower p-values (Fig. 5), confirming that our approach can predict expression patterns over large evolutionary distances. For some clades, the available data was very uneven on the two sides of the taxonomic split. For example, at the level of eumetazoa, only two species with few tissues were available for cnidarians, whereas most bilaterian species had many tissues measured. Thus, expression distances were higher when mapping from cnidarians to bilaterians than the other way round. Interestingly, the median divergence between animals and the outgroup choanoflagellates was comparable to the median divergence between major animal clades, e.g. bilateria.

When we chose an expression distance cutoff of 0.25 to designate well-conserved genes, we found that 77% of all 1:1 orthologs could be mapped successfully between mouse and human (Fig. 6). For larger clades (like vertebrates), we computed for each OG the median of all pairwise expression distances between the subclades (in this example, tetrapods and fish). Between tetrapods and fish, we found that 55% of all OGs have an expression distance below 0.25. Between animals and the outgroup choanoflagellate, 32.7% of all 1:1 orthologs showed conserved expression, which is a significant increase over the 25% expected when 1:1 orthologs behave like background genes (p-value of one-sided binomial test: 3e-23). Thus, mapping tissue-specific gene expression revealed expression programs conserved for 1 billion years. As the median expression similarities were negatively influenced by datasets of low quality or small size, we also computed the distributions of expression distances and the number of well-conserved OGs for the best dataset pair across each taxonomic split (Fig. S6 and Fig. S7).

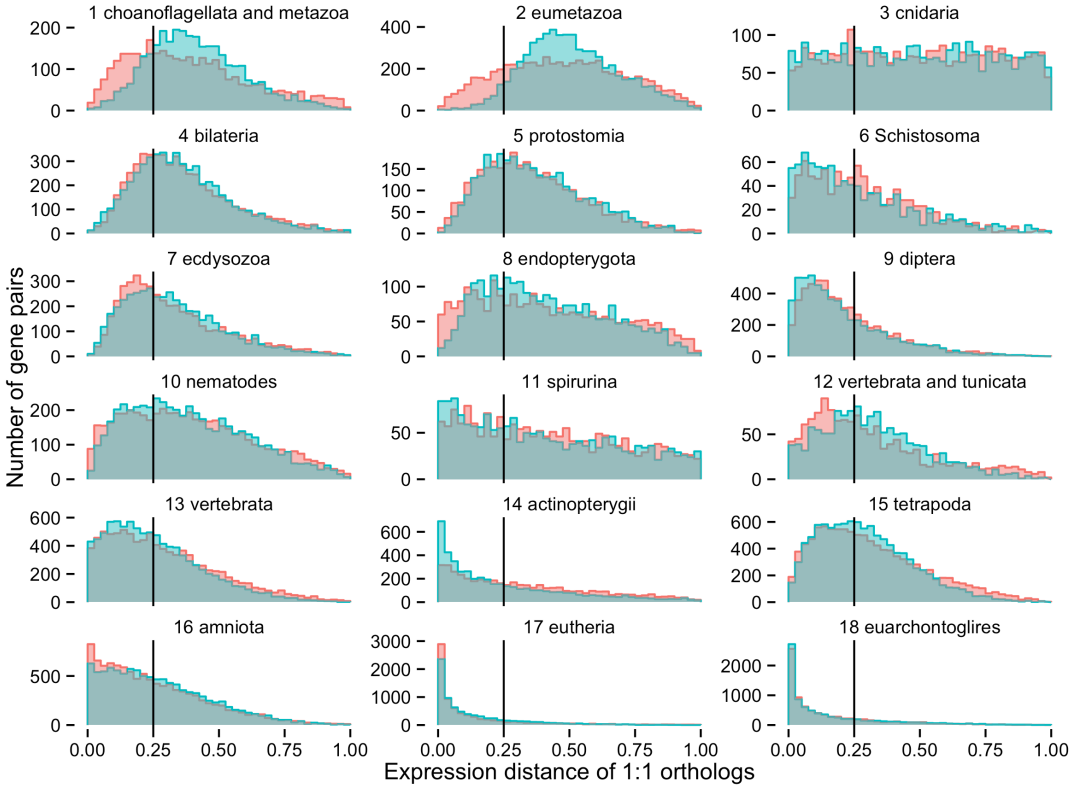


Fig. 5: Distribution of median conserved expression. For each clade, the distribution of expression distances of 1:1 orthologs is shown. Red and blue colors denote the direction of the mapping, either from the first subclade to the second or vice versa. The black bar corresponds to the expression distance cutoff of 0.25 (Fig. 6). When the mapping is successful, our mapping procedure yields virtual expression patterns of 1:1 orthologs that are very similar to the actual expression patterns, and the distribution of expression distances is skewed towards lower values. Our mapping procedure becomes less accurate over larger evolutionary distances, and the distribution of expression distances becomes less skewed. It becomes a uniform distribution when 1:1 orthologs cannot be mapped better than background gene pairs. Clades are numbered corresponding to the taxonomic tree in Fig. 6.

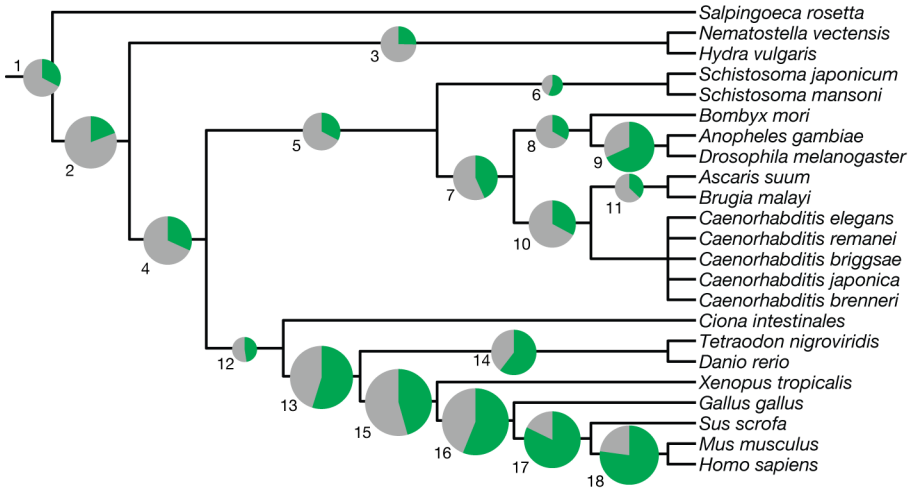


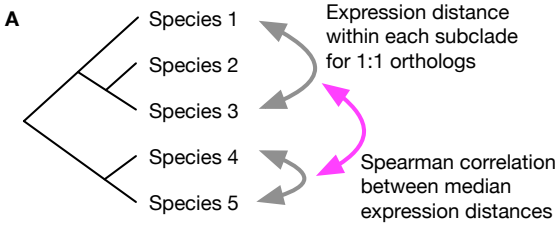
Fig. 6: Median conserved expression across animal clades. At each bifurcation, the pie chart denotes the median fraction of 1:1 orthologs with expression distances below 0.25. The area corresponds to the number of 1:1 orthologs across the taxonomic split. The numbers below the pie charts refer to the clade numbers in Fig. 5.

The conservation of conservation

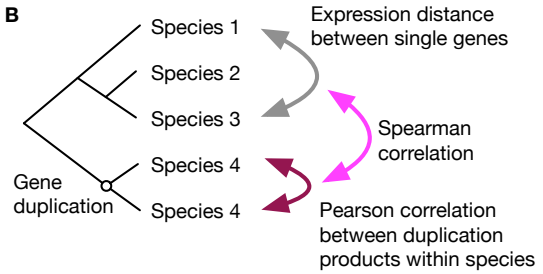
In the previous section, we showed that there is an enrichment in conserved expression programs across most taxonomic splits. Here, we analyzed to what extent the conservation of expression programs (i.e. the expression distance between family members) depends on the gene family. If the rate of expression divergence is a property of the gene family we expect a correlation between the expression similarities for each family in different clades. In other words, a gene that has a conserved expression pattern in one clade should also have a conserved expression pattern in another clade. For each taxonomic split with two or more species on either side of the split, we calculated the median expression distance per gene family within each of the two clades. Out of six taxonomic splits with more than one species on both sides, we found significant Spearman correlations (r_s) of median expression similarities for three splits (Fig. 7A): between tetrapods and fishes (r_s=0.18, #13 in Fig. 6), between protostomes and deuterostomes (r_s=0.14, #4), and between nematodes and insects (r_s=0.074, #7). Not significant were the splits involving cnidaria (#2), *Schistosoma* (#5) and spirurina (#10). Thus, expression distances were correlated across most taxonomic splits.

The previous analysis was only possible for a subset of the taxonomic splits in our body of data, due to the requirement of having more than one species on either side of the split. We therefore also analyzed the fate of duplicated genes. In this case, we tested whether duplication products are more similar if the non-duplicated members

of the gene family have low expression distances across the species outside the duplication event. Indeed, we found significant negative correlations between the median expression distance among the non-duplicated genes and the intra-species correlation of the duplicated genes (Fig. 7B). For example, duplicated genes in fish were more similar (i.e. has a higher correlation) when the corresponding tetrapod genes had more similar expression patterns (i.e. had a low expression distance): r_s=-0.11 for 1999 pairs of duplicated genes, corresponding to a p-value of 1e-7. Taken together, these two observations implied that for a significant fraction of genes, the rates of change in gene expression patterns were correlated between independently evolving clades.



#	clade	r_s	p-value	n
2	eumetazoa	0.04	0.03	2478
4	bilateria	0.14	0	4201
5	protostomia	0.05	0.24	629
7	ecdysozoa	0.07	1E-05	3467
10	rhabditida	0.04	0.13	1458
13	vertebrata	0.18	0	3391



#	clade with duplicated gene	clade with single genes	r_s	p-value	n
1	*Salpingocea*	eumetazoa	-0.14	0.08	164
4	deuterostomia	protostomia	-0.03	0.004	8618
5	*Schistosoma*	ecdysozoa	-0.11	0.02	440
5	ecdysozoa	*Schistosoma*	-0.10	0.02	560
7	endopterygota	rhabditida	-0.07	0.02	1120
12	*Ciona*	vertebrata	-0.39	0.03	29
13	actinopterygii	tetrapoda	-0.12	1E-07	1999

Fig. 7. Correlations between expression conservation rates. A For 1:1 orthologs, the expression distance across taxonomic splits was compared. In the dataset, there were only six splits with at least two species on both sides of the split. For example, when genes were similar within tetrapods, they also tended to be similar within fishes. **B** The rate at which gene duplication products diverge was negatively correlated with the expression distance among single-copy genes in related species. Only correlations with p-values below 0.1 are shown in this table. (# – Number of clade in Fig. 6, n – count of 1:1 orthologs [A] or duplicated genes [B])

Evolution of the beta catenin protein family

We have chosen the beta catenin protein family (Peifer et al. 1992) as an example to illustrate the implications of our work. Beta catenin proteins are involved in regulating cell adhesion and gene transcription through the Wnt signaling pathway. Ancestrally, there was a single beta catenin protein, which duplicated independently in the nematode and vertebrate lineages (Zhao et al. 2011). Hence, *Drosophila*, *Anopheles* and *Schistosoma* have only one beta-catenin, armadillo. This protein was

similar in its expression patterns with both the vertebrate and nematode beta-catenins (Fig. 8), which is indicative of their functional similarities (White et al. 1998). In vertebrates, two forms exist: beta-catenin and plakoglobin. These two proteins have largely overlapping functions (Swope et al. 2013) and consequently, their observed expression distance was very low. In nematodes, the outcome of the repeated gene duplications (Liu et al. 2008; Korswagen et al. 2000; Natarajan et al. 2001) is very different: three of the duplication products (hmp-2, wrm-1, and sys-1) are very similar to each other in their expression patterns, which can be explained by their cooperation in in the non-canonical Wnt signaling pathway and the SYS pathway (Kidd et al. 2005). These three proteins showed a significant dissimilarity in their expression patterns compared to bar-1. In contrast to them, bar-1 is part of a canonical Wnt signaling pathway (Kidd et al. 2005). We also observed that bar-1 had a low expression distance to the two vertebrate beta catenins, while hmp-2, wrm-1, and sys-1 showed significant dissimilarity with plakoglobin. This example illustrates that our method is able to uncover patterns of expression similarity and divergence both between closely related species and across large evolutionary distances.

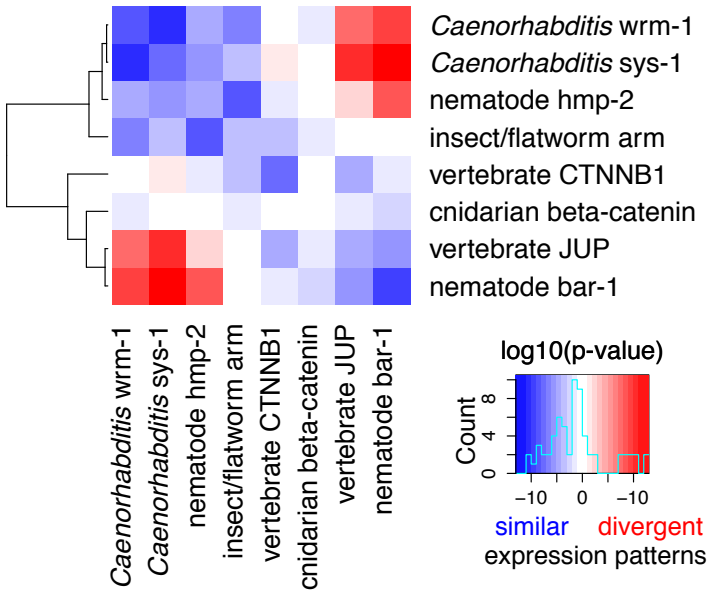


Fig. 8. Expression similarity and divergence in the beta catenins. Based on the expression divergence scores of the individual proteins (Fig. S8), we computed p-values for the expression similarity and divergence of groups of proteins (see Methods): Blue cells indicate that the proteins have low expression divergence scores, and red cells that they have high expression divergence scores. Profiles are clustered based on their Pearson correlation. The unduplicated beta catenins from insects, flatworms and cnidarians a similar to all other protein groups, while functional and expression divergence has occurred among the nematodes.

Resolving the fate of gene duplication products

The above results suggested that our expression distance measure could be used to compare functions of gene duplication products (i.e. in-paralogs) across species (i.e. with their respective orthologs). To this end, we created an additional expression distance metric that combines measures for expression similarity and dissimilarity, which we term "expression divergence score." This allowed us to also test if two genes have significantly diverging expression patterns. As above, we used the p-value for the null hypothesis that the genes are not related to each other (p_b) to quantify expression similarity. To measure expression dissimilarity, we used the p-value for the null hypothesis that considered genes are in fact 1:1 orthologs (p_o). We then combined the two p-values into an expression divergence score *E*:

$$E = \begin{cases} -\log_{10} p_o & p_o \leq p_b \\ \log_{10} p_b & p_o > p_b \end{cases}$$

Thus, the expression divergence score *E* is negative for similar, and positive for dissimilar gene pairs. Considering gene duplications, we computed divergence scores for both duplication products (Fig. 9). Using $\log_{10}(0.25)$ as a cutoff, we divided pairs of duplication products into three categories: (a) both genes had conserved expression patterns (2226 pairs of duplication products), (b) both genes had diverging expression patterns (911 pairs) and (c) only one of the duplication products had a diverging expression pattern, while the other one was conserved (1206 pairs). For each duplication event we also computed the expression conservation of the respective non-duplicated orthologs among each other (Fig. S9). It turned out that, when expression distances among the non-duplicated orthologs were small, the respective duplication products were more likely to be both conserved (Fig. 10), supporting again that purifying selection acts across large phylogenetic distances.

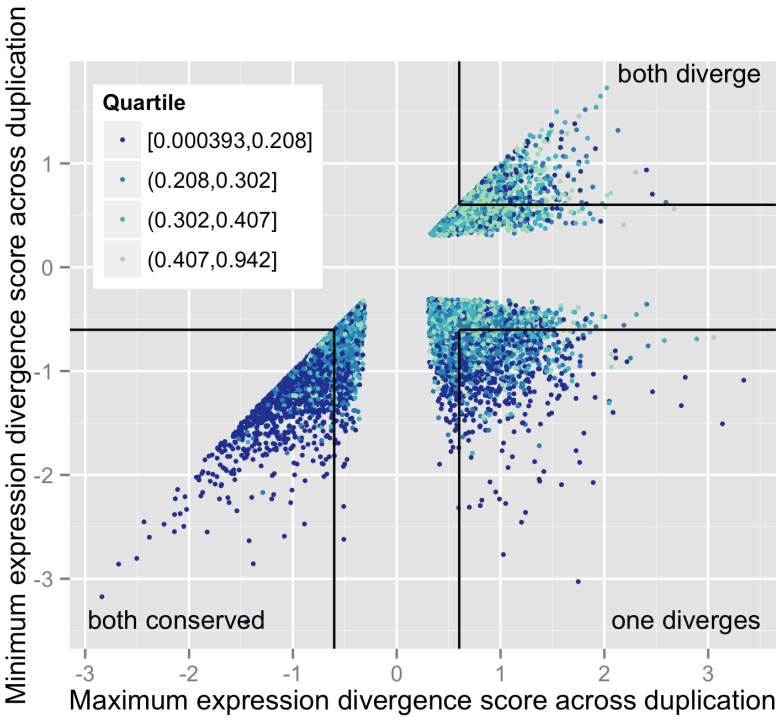


Fig. 9. Expression divergence scores of duplication products. For each duplication event, the expression divergence score of the duplication products to the non-duplicated genes was computed. For each pair of duplication products, the expression divergence scores were sorted. Thus, in the lower left quadrant, both duplication products had conserved expression patterns. In the upper right, both duplication products had diverging expression patterns and in the lower right, the outcome was mixed. Black lines denote an expression divergence score cutoff of 0.25.

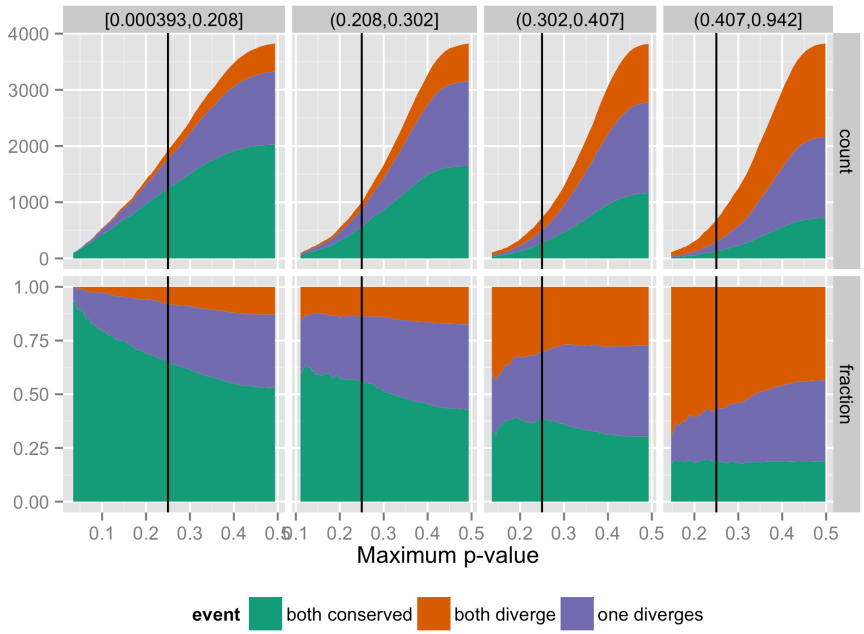


Fig. 10. Counts of duplication outcomes. Duplication events were grouped into quartiles according to the expression distance among the non-duplicated genes (Fig. S9). For each quartile, counts and fractions of the different outcomes are shown as the p-value cutoff is varied. This maximum p-value corresponds to a minimum of the absolute value of the expression divergence score, e.g. $|E| > \log_{10}(0.25)$ for the black lines.

Functional implications of diverging expression patterns in duplication products

In order to independently validate the functional implications of the observed conservation of expression programs between duplication products, we utilized protein–protein interaction (PPI) data from the STRING database (Franceschini et al. 2013). In particular, we investigated whether duplication products with expression programs more closely resembling non-duplicated orthologs are also functionally more related to the non-duplicated genes. Thus, for each pair of duplicated genes we designated the gene with the lower expression distance to non-duplicated orthologs as “less divergent” and the gene with the greater expression distance as “more divergent.”

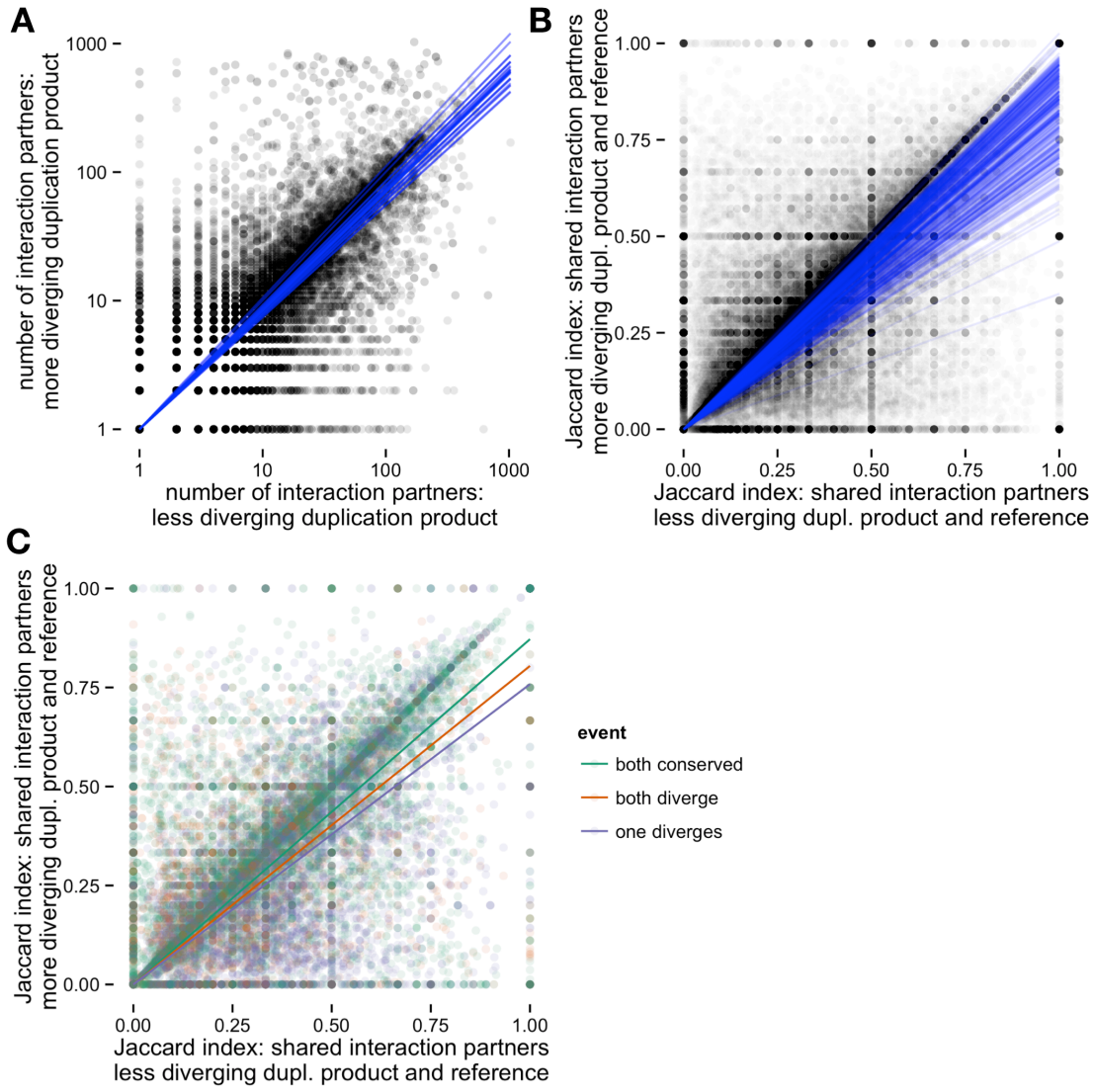


Fig. 11. Differences in the protein interaction network. (A) For all pairs of duplication products, we determined the number of interaction partners. Each dot corresponds to a pair of duplication products. Blue lines show linear fits for each dataset. **(B)** Here, each dot corresponds to the combination of a reference species (where the gene has not been duplicated) and a pair of duplication products. Blue lines correspond to linear fits for each combination of reference species and dataset. **(C)** The subset of duplication products is shown for which the outcome of the duplication could be determined. The difference between the two duplication products is largest for when only one of the duplication products diverges in its expression pattern. Here, the complete STRING network with confidence cutoff 0.5 is used. For an analysis of other networks and cutoffs, see Fig. S10, Fig. S11, and Fig. S12.

First, we found that less divergent genes had significantly more interaction partners than more divergent genes in 15 out of 25 datasets with available STRING data (using a p-value cutoff of 0.05 for one-sided Wilcoxon signed-rank tests, Fig. 11A and Fig. S10). From this data, it remained unclear if the less divergent protein gained interaction partners, or if the more divergent protein lost interaction partners.

However, the latter hypothesis seemed more parsimonious to us: interaction partners are often tissue specific. Thus, if the diverged protein got expressed in different tissues it likely lost some of its former interaction partners. To further corroborate this notion, we compared the interaction partners of the duplication products with the interaction partners of the respective non-duplicated genes.

We mapped PPI across species by counting how many OGs were shared between the interaction partners of the duplication product and the reference protein in a second species. We then calculated the Jaccard indices for the shared interaction partners between the reference protein and either of the duplication products (Fig. 11B). We found that for 40.5% of 395 dataset–species pairs, less divergent genes had a significantly higher Jaccard index compared to the more divergent genes (Fig. S11), while there were no dataset–species pairs where the more divergent proteins had significantly higher Jaccard indices. Furthermore, when we distinguished the possible outcomes discussed above (both duplication products have conserved expression patterns, one diverges, or both diverge), major differences only occurred for the case of one gene diverging (Fig. 11C and Fig. S12): in this case, less divergent genes had higher Jaccard indices in 40.0% of 175 dataset–species pairs, and all other outcomes were much less prevalent (<7%). (To be able to compare the p-values between the different outcomes, the same number of duplication products were used for the significance tests by sub-sampling 100 times.) This observation is consistent with earlier findings about the tissue specificity of protein complexes (Börnigen et al. 2013) and strengthens the notion that the duplication product with the more divergent expression pattern lost previous interactions and acquired novel interaction partners.

Expression divergence depends on protein function

In order to evaluate the extent to which the conservation of expression patterns after gene duplications depends on the function of gene products, we assigned protein classes according to the PANTHER database (Mi et al. 2013) to all OGs. We then determined the significance of the enrichment of gene duplication events where both genes were conserved (bottom left corner of Fig. 9) versus cases where one gene diverged (bottom right corner of Fig. 9, see Methods for more details). Extracellular matrix proteins, transcription factors, and receptors were significantly enriched for duplication events with conserved expression patterns (Fig. 12, top). Conversely, chaperones, calcium-binding proteins and transfer/carrier proteins are

enriched for duplication events with diverging expression patterns (Fig. 12, bottom), suggesting that regulatory proteins tend to be more conserved, whereas proteins executing tissue-specific functions are more likely to diverge.

Fig. 12. Gene duplication outcomes differ between protein classes. For each species and protein class, we analyzed duplications where one or both of the duplication products were conserved. **Left:** fraction of duplications where both duplication products had conserved expression patterns. **Right:** fractions normalized according to the species-wise average. Black bars correspond to the weighted average across all species. One-sided p-values have been calculated using a Poisson binomial distribution for each protein class, comparing the actual number of events to the expected number based on the species-wise averages. “Not assigned” marks OGs that could be mapped to PANTHER, but for which no protein class had been assigned in the PANTHER database. “Not mapped” stands for OGs for which no member could be mapped to the PANTHER database. Using an iterative method (see Methods) to remove multiple annotations per OG, we find a slightly different set of significant protein classes (Fig. S13). For example, nucleic acid binding proteins encompass transcription factors and are therefore enriched.

Discussion

The presented analysis established and benchmarked a new method, and provided four biological conclusions: there is widespread conservation of expression regulation across very large evolutionary distances; independently evolving members of the same gene family have significant correlations in their speed of divergences; expression divergence can be used to grade the functional conservation of gene duplication products; and neofunctionalization of gene duplication products is dependent on gene function.

In particular, we have shown that tissue-specific gene expression can be predicted across large evolutionary distances, even in the absence of apparent similarities between the species' tissues. Our approach can be rationalized as follows: we assume that evolution conserves the co-expression of functionally related genes, both on the level of homologous cell types and on the level of functional modules that occur in unrelated tissues. Our analysis demonstrated that the expression patterns of such conserved gene modules can be predicted across species using 1:1 orthologs as "anchors." This approach worked despite the fact that the tissues themselves are only conserved within smaller clades. Control of gene expression by transcription factors, miRNAs and other factors is known to turn over rather quickly (Odom et al. 2007; Bradley et al. 2010; Berezikov 2011). Most probably, functional dependencies between genes lead to shared expression patterns over large evolutionary distances. Further research will be needed to reveal which expression similarities between tissues are caused by homology and which are caused by convergent evolution.

When we applied the concept of looking for correlations between orthologs across species to an existing dataset (Brawand et al. 2011), we found that many of the reported lineage-specific expression shifts only changed the absolute expression levels, while the relative expression patterns remained conserved (Fig. S14). This suggests that further studies could combine approaches that test absolute and relative expression patterns to identify truly novel expression patterns. We investigated products of gene duplication events and found that they seem to have the ability to "opt out" of such gene expression modules to acquire new functions. Such events suggest unidirectional dependences: whereas the duplicated gene does not need (all of) its ancient interaction partners, the partners seem to need the

duplicated gene and, thus, one of the two remained in the respective expression module. A more detailed analysis, including the divergence on the level of the protein sequence and the mode of natural selection (such as positive or neutral selection) may lead to more connections between expression divergence, protein function and sequence evolution.

Methods

Import of expression data

Datasets were obtained either from repositories like ArrayExpress and GEO, from supplementary materials or the respective websites of the resources. Expression profiles were then mapped to our set of genes by one of the following methods (see Table S1): If possible, genes were mapped by given identifiers, such as Affymetrix, Ensembl or WormBase identifiers. If identifiers could not be used for microarrays, we mapped probe sequences to transcripts using exonerate (Slater & Birney 2005), allowing for up to three mismatches and discarding probes that mapped to multiple genes. In the case of RNA-seq data without matching identifiers, we mapped reads to annotated transcripts using tophat2 and cufflinks 2.1.1 (Kim et al. 2013; Trapnell et al. 2010) and used the resulting FPKM counts.

Normalization of expression data

In initial small-scale tests, we tested several normalization methods (Liao & Zhang 2006; Piasecka, Robinson-Rechavi, et al. 2012b), and settled on a *z*-like normalization of expression vectors **x**, which corresponds to the Euclidean normalization of **x** minus its median value. Therefore, we did not look for conserved expression abundance, but rather for conserved relative expression across tissues. Normalizing each gene's expression individually also avoided technical concerns regarding the comparability of absolute expression values between genes. RNA-seq data, e.g. the *Drosophila* modENCODE dataset, contained zeros, which were of course not suitable for logarithmic analysis. For these datasets, we determined the expression value of the 1/1000th quantile of all genes with non-zero expression. All expression values were incremented by this value.

Tissue correlations between species

P-values for tissue correlations were calculated analytically. We performed tests with shuffling of genes to confirm that the analytical p-values correspond to empirical p-values.

Mapping of tissue expression patterns

For each pair of datasets, individual linear models were trained for each tissue of the target species, using the tissues of the source species as input. (Note that due to the normalization, one tissue is redundant and therefore left out. This also implies that the coefficients of the linear model are not directly interpretable.) The set of 1:1 orthologs between the two species was used as a training set. When there were multiple probes per gene, all combinations of probes were used for training. When there are many tissues in the source species, but few 1:1 orthologs, there is the danger of over-fitting. We therefore allowed only one predictor (i.e. one tissue from the source species) per 15 samples (1:1 orthologs) (Babyak 2004). For each pair of species, the safe number of predictors was calculated. If there were too many tissues, we combined tissues using *k*-means clustering and used the centers of the clusters as predictors. This situation only occurred for six out of 992 dataset pairs. The trained models are then applied to all genes of the source species, yielding corresponding predicted expression patterns in the target species. Since 1:1 orthologs are used for training, we used predictions from a 10-fold cross-validation for these genes.

Computation of expression distances

For each pair of datasets, we computed a matrix of predicted expression patterns of all genes from the source species. We then calculated the weighted Pearson correlations between the mapped expression patterns and the actual expression patterns of the target species' genes. Weights on the tissues were calculated using the Gerstein-Sonnhammer-Chothia (GSC) weighting scheme to reduce the effect of uneven coverage of different anatomical regions (Gerstein et al. 1994). For example, in the mouse tissue datasets, there are many different brain tissues. Given the matrix of all weighted Pearson correlations, we then calculated expression distances like p-values, i.e. by determining the fraction of unrelated genes that have the same or higher correlation. For technical reasons, we sampled one million pairs of background genes, such that the lowest possible expression distance is 1e-6.

As mentioned in the Results section, there is a strong correlation between the raw expression distances and the number of genes in the target species. This strong correlation indicated that predictions were biased towards the average target gene (i.e. the average expression profile of all genes considered in the target species), which in turn was similar to many target genes. As a consequence, these "close-to-

average" target genes had higher correlations with mapped source genes, and thus seemed more conserved. To counter this effect, target genes are split into ten bins according to the number of co-expressed genes in the target species. Thus, there exist ten conversion functions from weighted Pearson correlation to an uncorrected expression distance. For a given pair of genes, the final expression distance is interpolated from the two adjacent bins. We determined the number of co-expressed genes for each target gene as follows: we first computed all pairwise correlations among the target genes of the training set. Then, we determined the correlation cutoff corresponding to the top 10%, and counted for each gene how many other target genes were among the global top 10% correlations.

Quantifying expression similarity and divergence for groups of genes

Similarly to the definition of the expression divergence score E, we take two p-values into account for each pair of datasets: the p-value for the null hypothesis that the genes are not related to each other (p_b) and the p-value for the null hypothesis that considered genes are in fact 1:1 orthologs (p_o). Given two groups of proteins, we then consider all interspecies combinations of datasets and compute paired Wilcoxon signed-rank tests to determine if the expression patterns are significantly similar ($p_b < p_o$) or divergent ($p_b > p_o$). For two groups of genes, we then report the lower p-value (Fig. 8).

Analysis of protein classes

Protein classes were obtained from the PANTHER 9.0 database (Mi et al. 2013) and mapped to OGs. For each combination of species and protein class, we then determined the fraction of duplication products where both products have conserved expression patterns, relative to the number of duplication products where at least one of the duplication products keeps a conserved expression pattern. For each species, there is thus a background frequency of duplication products that are both conserved. For each class, we then determined p-values for both over- and under-representation using a Poisson binomial distribution. This distribution is computed from the background frequencies and the number of duplications of the relevant class per species. As the protein classes overlap (e.g. kinases are also transferases), we also used an iterative method to determine p-values: after each iteration, all OGs that were annotated with the protein class with the most significant p-value were removed from the analysis (while keeping the

background frequencies constant). Thus, the new p-value for nucleic acids binding proteins excludes transcription factors, as these had a lower p-value (Fig. S13).

Acknowledgements

The authors thank Anthony A. Hyman and Vineeth Surendranath for helpful discussions.

Funding

MK is funded by the Deutsche Forschungsgemeinschaft (DFG KU 2796/2-1). AB receives funding from the Deutsche Forschungsgemeinschaft (DFG CRC 680).

Author Contributions

AB and MK conceived the study, planned the analyses and wrote the paper. MK conducted all analyses.

Competing Interests

The authors declare that there are no competing interests.

References

Babyak, M.A., 2004. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic medicine*, 66(3), pp.411–421.

Baker, D.A. et al., 2011. A comprehensive gene expression atlas of sex- and tissue-specificity in the malaria vector, Anopheles gambiae. *BMC Genomics*, 12, p.296. doi:10.1186/1471-2164-12-296.

Benjamini, Y. & Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), pp.289–300. doi:10.2307/2346101.

Berezikov, E., 2011. Evolution of microRNA diversity and regulation in animals. *Nature Reviews Genetics*, 12(12), pp.846–860. doi:10.1038/nrg3079.

Börnigen, D. et al., 2013. Concordance of gene expression in human protein complexes reveals tissue specificity and pathology. *Nucleic acids research*, 41(18), p.e171. doi:10.1093/nar/gkt661.

Bradley, R.K. et al., 2010. Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. *PLoS biology*, 8(3), p.e1000343. doi:10.1371/journal.pbio.1000343.

Brawand, D. et al., 2011. The evolution of gene expression levels in mammalian

organs. *Nature*, 478(7369), pp.343–348. doi:10.1038/nature10532.

Chan, E.T. et al., 2009. Conservation of core gene expression in vertebrate tissues. *Journal of biology*, 8(3), p.33. doi:10.1186/jbiol130.

Chikina, M.D. & Troyanskaya, O.G., 2011. Accurate Quantification of Functional Analogy among Close Homologs. *PLoS computational biology*, 7(2), p.e1001074. doi:10.1371/journal.pcbi.1001074.

Chikina, M.D. et al., 2009. Global prediction of tissue-specific gene expression and context-dependent gene networks in Caenorhabditis elegans. *PLoS computational biology*, 5(6), p.e1000417. doi:10.1371/journal.pcbi.1000417.

Chung, H. et al., 2009. Characterization of Drosophila melanogaster cytochrome P450 genes. *Proceedings of the National Academy of Sciences of the United States of America*, 106(14), pp.5731–5736. doi:10.1073/pnas.0812141106.

Dissanayake, S.N. et al., 2006. angaGEDUCI: Anopheles gambiae gene expression database with integrated comparative algorithms for identifying conserved DNA motifs in promoter sequences. *BMC Genomics*, 7, p.116. doi:10.1186/1471-2164-7-116.

Domazet-Loso, T. & Tautz, D., 2010. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature*, 468(7325), pp.815–818. doi:10.1038/nature09632.

Fairclough, S.R. et al., 2013. Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate Salpingoeca rosetta. *Genome biology*, 14(2), p.R15. doi:10.1186/gb-2013-14-2-r15.

Fitzpatrick, J.M. et al., 2009. Anti-schistosomal intervention targets identified by lifecycle transcriptomic analyses. *PLoS neglected tropical diseases*, 3(11), p.e543. doi:10.1371/journal.pntd.0000543.

Franceschini, A. et al., 2013. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*, 41(Database issue), pp.D808–15. doi:10.1093/nar/gks1094.

Freeman, T.C. et al., 2012. A gene expression atlas of the domestic pig. *BMC Biology*, 10, p.90. doi:10.1186/1741-7007-10-90.

Gerstein, M.B. et al., 2014. Comparative analysis of the transcriptome across distant species. *Nature*, 512(7515), pp.445–448. doi:doi:10.1038/nature13424.

Gerstein, M.B., Sonnhammer, E.L.L. & Chothia, C., 1994. Volume changes in protein evolution. *Journal of Molecular Biology*, 236(4), pp.1067–1078.

Gobert, G.N. et al., 2009. Developmental gene expression profiles of the human pathogen Schistosoma japonicum. *BMC Genomics*, 10, p.128. doi:10.1186/1471-2164-10-128.

Goltsev, Y. et al., 2009. Developmental and evolutionary basis for drought tolerance of the Anopheles gambiae embryo. *Developmental biology*, 330(2), pp.462–470. doi:10.1016/j.ydbio.2009.02.038.

Gu, X. & Su, Z., 2007. Tissue-driven hypothesis of genomic evolution and sequence-expression correlations. *Proceedings of the National Academy of Sciences of the United States of America*, 104(8), pp.2779–2784. doi:10.1073/pnas.0610797104.

Hemmrich, G. et al., 2012. Molecular signatures of the three stem cell lineages in hydra and the emergence of stem cell function at the base of multicellularity. *Molecular Biology and Evolution*, 29(11), pp.3267–3280. doi:10.1093/molbev/mss134.

Irie, N. & Kuratani, S., 2011. Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. *Nature communications*, 2, p.248. doi:10.1038/ncomms1248.

Kidd, A.R. et al., 2005. A beta-catenin identified by functional rather than sequence criteria and its role in Wnt/MAPK signaling. *Cell*, 121(5), pp.761–772. doi:10.1016/j.cell.2005.03.029.

Kim, D. et al., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14(4), p.R36. doi:10.1186/gb-2013-14-4-r36.

Korswagen, H.C., Herman, M.A. & Clevers, H.C., 2000. Distinct beta-catenins mediate adhesion and signalling functions in C. elegans. *Nature*, 406(6795), pp.527–532. doi:10.1038/35020099.

Lees, J.G. et al., 2014. Gene3D: Multi-domain annotations for protein sequence and comparative genome analysis. *Nucleic acids research*, 42(1), pp.D240–5. doi:10.1093/nar/gkt1205.

Levin, M. et al., 2012. Developmental milestones punctuate gene expression in the Caenorhabditis embryo. *Developmental cell*, 22(5), pp.1101–1108. doi:10.1016/j.devcel.2012.04.004.

Liao, B.-Y. & Zhang, J., 2006. Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Molecular Biology and Evolution*, 23(3), pp.530–540. doi:10.1093/molbev/msj054.

Liu, J. et al., 2008. The C. elegans SYS-1 protein is a bona fide beta-catenin. *Developmental cell*, 14(5), pp.751–761. doi:10.1016/j.devcel.2008.02.015.

Lukk, M. et al., 2010. A global map of human gene expression. *Nature Biotechnology*, 28(4), pp.322–324. doi:10.1038/nbt0410-322.

McGary, K.L. et al., 2010. Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proceedings of the National Academy of Sciences of the United States of America*, 107(14), pp.6544–6549. doi:10.1073/pnas.0910200107.

Mi, H., Muruganujan, A. & Thomas, P.D., 2013. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic acids research*, 41(Database issue), pp.D377–86. doi:10.1093/nar/gks1118.

Natarajan, L., Witwer, N.E. & Eisenmann, D.M., 2001. The divergent Caenorhabditis elegans beta-catenin proteins BAR-1, WRM-1 and HMP-2 make distinct protein interactions but retain functional redundancy in vivo. *Genetics*, 159(1), pp.159–172.

Nawaratna, S.S.K. et al., 2011. Gene Atlasing of digestive and reproductive tissues in Schistosoma mansoni. *PLoS neglected tropical diseases*, 5(4), p.e1043. doi:10.1371/journal.pntd.0001043.

Niknejad, A. et al., 2012. vHOG, a multispecies vertebrate ontology of homologous organs groups. *Bioinformatics (Oxford, England)*, 28(7), pp.1017–1020. doi:10.1093/bioinformatics/bts048.

Odom, D.T. et al., 2007. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nature Genetics*, 39(6), pp.730–732. doi:10.1038/ng2047.

Peifer, M. et al., 1992. The vertebrate adhesive junction proteins beta-catenin and plakoglobin and the Drosophila segment polarity gene armadillo form a multigene family with similar properties. *The Journal of cell biology*, 118(3), pp.681–691.

Piasecka, B., Kutalik, Z., et al., 2012a. Comparative modular analysis of gene expression in vertebrate organs. *BMC Genomics*, 13, p.124. doi:10.1186/1471-2164-13-124.

Piasecka, B., Robinson-Rechavi, M. & Bergmann, S., 2012b. Correcting for the bias due to expression specificity improves the estimation of constrained evolution of expression between mouse and human. *Bioinformatics (Oxford, England)*, 28(14), pp.1865–1872. doi:10.1093/bioinformatics/bts266.

Powell, S. et al., 2014. eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic acids research*, 42(1), pp.D231–9. doi:10.1093/nar/gkt1253.

Robinson, S.W. et al., 2013. FlyAtlas: database of gene expression in the tissues of Drosophila melanogaster. *Nucleic acids research*, 41(Database issue), pp.D744–50. doi:10.1093/nar/gks1141.

Seok, J. et al., 2013. Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proceedings of the National Academy of Sciences of the United States of America*, 110(9), pp.3507–3512. Available at: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=23401516&retmode=ref&cmd=prlinks.

Shoguchi, E. et al., 2011. Direct examination of chromosomal clustering of organ-specific genes in the chordate Ciona intestinalis. *Genesis (New York, N.Y. : 2000)*, 49(8), pp.662–672. doi:10.1002/dvg.20730.

Shubin, N., Tabin, C. & Carroll, S., 2009. Deep homology and the origins of evolutionary novelty. *Nature*, 457(7231), pp.818–823. doi:10.1038/nature07891.

Silver, D.H., Levin, M. & Yanai, I., 2012. Identifying functional links between genes by evolutionary transcriptomics. *Molecular BioSystems*, 8(10), pp.2585–2592. doi:10.1039/c2mb25054c.

Slater, G.S.C. & Birney, E., 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6, p.31. doi:10.1186/1471-2105-6-31.

Spencer, W.C. et al., 2011. A spatial and temporal map of C. elegans gene expression. *Genome research*, 21(2), pp.325–341. doi:10.1101/gr.114595.110.

St Pierre, S.E. et al., 2014. FlyBase 102--advanced approaches to interrogating FlyBase. *Nucleic acids research*, 42(Database issue), pp.D780–8. doi:10.1093/nar/gkt1092.

Strausfeld, N.J. & Hirth, F., 2013. Deep Homology of Arthropod Central Complex and Vertebrate Basal Ganglia. *Science (New York, NY)*.

Stuart, J.M., 2003. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science (New York, NY)*, 302(5643), pp.249–255. doi:10.1126/science.1087447.

Su, A.I. et al., 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16), pp.6062–6067. doi:10.1073/pnas.0400782101.

Swope, D., Li, J. & Radice, G.L., 2013. Beyond cell adhesion: the role of armadillo proteins in the heart. *Cellular signalling*, 25(1), pp.93–100. doi:10.1016/j.cellsig.2012.09.025.

Thomas, P.D., 2010. GIGA: a simple, efficient algorithm for gene tree inference in the genomic age. *BMC Bioinformatics*, 11, p.312. doi:10.1186/1471-2105-11-312.

Thomas, P.D. et al., 2012. On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report. *PLoS computational biology*, 8(2), p.e1002386. doi:10.1371/journal.pcbi.1002386.

Tomer, R. et al., 2010. Profiling by image registration reveals common origin of annelid mushroom bodies and vertebrate pallium. *Cell*, 142(5), pp.800–809. doi:10.1016/j.cell.2010.07.043.

Trapnell, C. et al., 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5), pp.511–515. doi:10.1038/nbt.1621.

Tulin, S. et al., 2013. A quantitative reference transcriptome for Nematostella vectensis early embryonic development: a pipeline for de novo assembly in emerging model systems. *EvoDevo*, 4(1), p.16. doi:10.1186/2041-9139-4-16.

Wang, Z. et al., 2013. Gene expression analysis distinguishes tissue-specific and gender-related functions among adult Ascaris suum tissues. *Molecular genetics and genomics : MGG*, 288(5-6), pp.243–260. doi:10.1007/s00438-013-0743-y.

White, P., Aberle, H. & Vincent, J.P., 1998. Signaling and adhesion activities of mammalian beta-catenin and plakoglobin in Drosophila. *The Journal of cell biology*, 140(1), pp.183–195.

Winter, E.E., Goodstadt, L. & Ponting, C.P., 2004. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome research*, 14(1), pp.54–61. doi:10.1101/gr.1924004.

Xia, Q. et al., 2007. Microarray-based gene expression profiles in multiple tissues of the domesticated silkworm, Bombyx mori. *Genome biology*, 8(8), p.R162. doi:10.1186/gb-2007-8-8-r162.

Yanai, I. et al., 2011. Mapping gene expression in two Xenopus species: evolutionary constraints and developmental flexibility. *Developmental cell*, 20(4), pp.483–496. doi:10.1016/j.devcel.2011.03.015.

Zhao, Z.-M., Reynolds, A.B. & Gaucher, E.A., 2011. The evolutionary history of the catenin gene family during metazoan evolution. *BMC Evolutionary Biology*, 11, p.198. doi:10.1186/1471-2148-11-198.